Assignment 6: GLMs (Linear Regressios, ANOVA, & t-tests)

Yared S. Asfaw

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on generalized linear models.

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Fay_A06_GLMs.Rmd") prior to submission.

The completed exercise is due on Monday, February 28 at 7:00 pm.

Set up your session

1. Set up your session. Check your working directory. Load the tidyverse, agricolae and other needed packages. Import the *raw* NTL-LTER raw data file for chemistry/physics (`NTL-LTER_Lake_ChemistryPhysics_Raw.csv`). Set date columns to date objects.

```
# Checking the working directory
getwd()
```

```
## [1] "E:/EDA/Environmental_Data_Analytics_2022"
```

```
# Loading important packages
library(tidyverse)
library(agricolae)
library(dplyr)
library(lubridate)
library(ggplot2)
#install.packages("viridisLite")
library(viridis)
library(RColorBrewer)
```

```
# Importing the dataset
ntl_lter <- read.csv(file = "./Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv",
                     stringsAsFactors = TRUE)
head(ntl_lter)
```

```
##   lakeid   lakename year4 daynum sampledate depth temperature_C dissolvedOxygen
## 1      L Paul Lake  1984    148    5/27/84  0.00          14.5             9.5
## 2      L Paul Lake  1984    148    5/27/84  0.25            NA              NA
## 3      L Paul Lake  1984    148    5/27/84  0.50            NA              NA
## 4      L Paul Lake  1984    148    5/27/84  0.75            NA              NA
## 5      L Paul Lake  1984    148    5/27/84  1.00          14.5             8.8
## 6      L Paul Lake  1984    148    5/27/84  1.50            NA              NA
##   irradianceWater irradianceDeck comments
## 1            1750           1620     <NA>
## 2            1550           1620     <NA>
## 3            1150           1620     <NA>
## 4             975           1620     <NA>
## 5             870           1620     <NA>
## 6             610           1620     <NA>
```

```r
# Formatting the date column to Date
head(ntl_lter$sampledate)
```

```
## [1] 5/27/84 5/27/84 5/27/84 5/27/84 5/27/84 5/27/84
## 1712 Levels: 10/1/07 10/1/93 10/12/02 10/14/88 10/15/87 10/15/88 ... 9/9/92
```

```r
ntl_lter$sampledate <- as.Date(ntl_lter$sampledate, format="%m/%d/%y")
```

2. Build a ggplot theme and set it as your default theme.

```r
# Building a ggplot theme
mytheme <- theme_classic(base_size = 12) +
  theme(axis.text = element_text(color = "black"), legend.position ="right")

# Setting up my_theme as default theme
theme_set(mytheme)
```

Simple regression

Our first research question is: Does mean lake temperature recorded during July change with depth across all lakes?

3. State the null and alternative hypotheses for this question: > Answer: H0: The mean lake temprature recorded during July doesn't change with depth (remain the same with) accross all lakes. Ha: At least one of the lakes' mean lake temprature recorded with depth during July differs

4. Wrangle your NTL-LTER dataset with a pipe function so that the records meet the following criteria:

- Only dates in July.
- Only the columns: `lakename`, `year4`, `daynum`, `depth`, `temperature_C`
- Only complete cases (i.e., remove NAs)

```r
selected_ntl_nter <- ntl_lter %>%
  filter(month(sampledate) == 7) %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  na.omit()
head(selected_ntl_nter)
```

```
##    lakename year4 daynum depth temperature_C
## 1 Paul Lake  1984    183   0.0          22.8
## 3 Paul Lake  1984    183   0.5          22.9
## 5 Paul Lake  1984    183   1.0          22.8
## 6 Paul Lake  1984    183   1.5          22.7
## 7 Paul Lake  1984    183   2.0          21.7
## 8 Paul Lake  1984    183   2.5          20.3
```

5. Visualize the relationship among the two continuous variables with a scatter plot of temperature by depth. Add a smoothed line showing the linear model, and limit temperature values from 0 to 35 °C. Make this plot look pretty and easy to read.

```
# Plotting the relationship between temperature and depth
ggplot(selected_ntl_nter, aes(x=depth, y=temperature_C)) +
  geom_point(aes(color=temperature_C)) +
  geom_smooth(color="blue", method = "lm") +
  ylim(label=c(0,35)) +
  ylab(expression("Temperature in "*degree*"C")) +
  xlab("Depth (m)")+
  scale_color_viridis(option = "magma") +
  ggtitle("Lake temperature by depth")
```
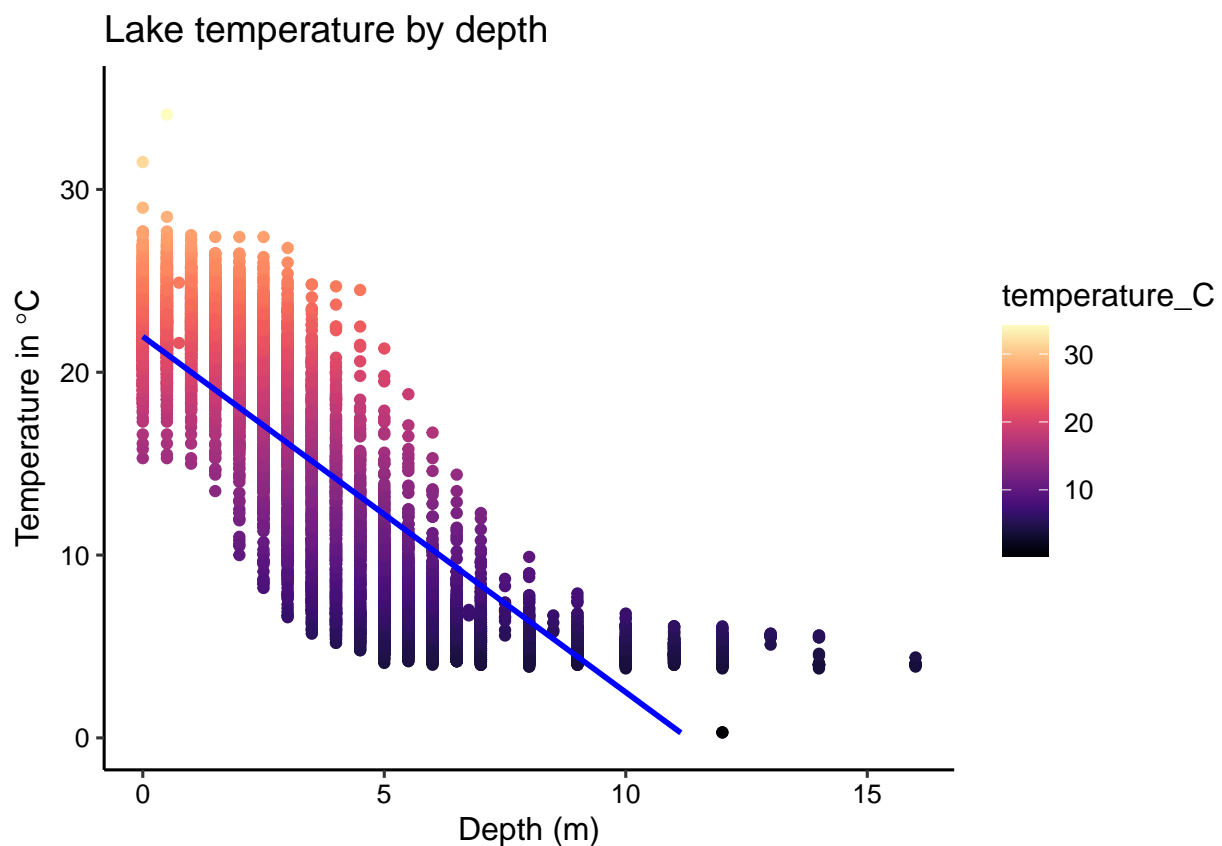

6. Interpret the figure. What does it suggest with regards to the response of temperature to depth? Do the distribution of points suggest about anything about the linearity of this trend?

3

Answer: As the depth of the lake increases, the temprature decreases. The way the points distributed show this trend significantly starting from the depth around 6-7m to the surface. Higher than 6-7m depth, there is a trend in the decrease in temperature but not aligned with the trend line. It seems a separate slight trend of temperature decrease as the depth increases starting from 10m.

7. Perform a linear regression to test the relationship and display the results

```
# Performing linear regression
selected_ntl_nter_reg <- lm(data = selected_ntl_nter, temperature_C ~ depth)
summary(selected_ntl_nter_reg)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = selected_ntl_nter)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5173 -3.0192  0.0633  2.9365 13.5834
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.95597    0.06792   323.3   <2e-16 ***
## depth       -1.94621    0.01174  -165.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.835 on 9726 degrees of freedom
## Multiple R-squared:  0.7387, Adjusted R-squared:  0.7387
## F-statistic: 2.75e+04 on 1 and 9726 DF,  p-value: < 2.2e-16
```

8. Interpret your model results in words. Include how much of the variability in temperature is explained by changes in depth, the degrees of freedom on which this finding is based, and the statistical significance of the result. Also mention how much temperature is predicted to change for every 1m change in depth.

Answer: Generally, depth and temperature are inversly related. As the depth increases by 1m, the temperature will decrease by 1.95 deree celisius. This model indicates that approximately 74% of the variablity in the lake temperature is explained by the changes in the depth of the lake. This finding is based on 9,726 degrees of freedom. The p-value is significantly less than the level of significance, thus, the result is statistically significant.

Multiple regression

Let's tackle a similar question from a different approach. Here, we want to explore what might the best set of predictors for lake temperature in July across the monitoring period at the North Temperate Lakes LTER.

9. Run an AIC to determine what set of explanatory variables (year4, daynum, depth) is best suited to predict temperature.

```
# Choose a model by AIC in a Stepwise Algorithm
tempAIC <- lm(data=selected_ntl_nter, temperature_C ~ year4 + daynum + depth)
step(tempAIC)
```

```
## Start:  AIC=26065.53
## temperature_C ~ year4 + daynum + depth
##
##           Df Sum of Sq     RSS   AIC
## <none>                  141687 26066
## - year4    1       101 141788 26070
## - daynum   1      1237 142924 26148
## - depth    1    404475 546161 39189
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = selected_ntl_nter)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##    -8.57556      0.01134      0.03978     -1.94644
```

10. Run a multiple regression on the recommended set of variables.

```
# Multiple regression on the recommended variables

Tempmodel <- lm(data = selected_ntl_nter, temperature_C ~ year4 + daynum + depth)
summary(Tempmodel)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = selected_ntl_nter)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6536 -3.0000  0.0902  2.9658 13.6123
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -8.575564   8.630715   -0.994  0.32044
## year4        0.011345   0.004299    2.639  0.00833 **
## daynum       0.039780   0.004317    9.215  < 2e-16 ***
## depth       -1.946437   0.011683 -166.611  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.817 on 9724 degrees of freedom
## Multiple R-squared:  0.7412, Adjusted R-squared:  0.7411
## F-statistic:  9283 on 3 and 9724 DF,  p-value: < 2.2e-16
```

11. What is the final set of explanatory variables that the AIC method suggests we use to predict temperature in our multiple regression? How much of the observed variance does this model explain? Is this an improvement over the model using only depth as the explanatory variable?

Answer: The set of variables suggested by the AIC method to predict temprature are year4, daynum and depth. This model explains 74.12% of the observed variance in the lake temperature which is almost the same percentage with the model that use depth only (i.e. 73.89%). Therefore, there is no improvement over the model that use depth only as explanatory variable.

Analysis of Variance

12. Now we want to see whether the different lakes have, on average, different temperatures in the month of July. Run an ANOVA test to complete this analysis. (No need to test assumptions of normality or similar variances.) Create two sets of models: one expressed as an ANOVA models and another expressed as a linear model (as done in our lessons).

```
# Checking whether the different lakes have different temperatures in July

# Defining null hypothesis and alternative hypothesis
## H0: All lakes have the same average temperatures in the month of July
## Ha: At least two lakes have different average temperatures in the month of July

# Model using ANOVA
lakeTemp_anova <- aov(data=selected_ntl_nter, temperature_C ~ lakename)
summary(lakeTemp_anova)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## lakename       8  21642  2705.2      50 <2e-16 ***
## Residuals   9719 525813    54.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Model using linear model
lakeTemp_lm <- lm(data=selected_ntl_nter, temperature_C ~ lakename)
summary(lakeTemp_lm)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = selected_ntl_nter)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.769  -6.614  -2.679   7.684  23.832
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              17.6664     0.6501  27.174  < 2e-16 ***
## lakenameCrampton Lake    -2.3145     0.7699  -3.006 0.002653 **
## lakenameEast Long Lake   -7.3987     0.6918 -10.695  < 2e-16 ***
## lakenameHummingbird Lake -6.8931     0.9429  -7.311 2.87e-13 ***
## lakenamePaul Lake        -3.8522     0.6656  -5.788 7.36e-09 ***
## lakenamePeter Lake       -4.3501     0.6645  -6.547 6.17e-11 ***
## lakenameTuesday Lake     -6.5972     0.6769  -9.746  < 2e-16 ***
## lakenameWard Lake        -3.2078     0.9429  -3.402 0.000672 ***
```

```
## lakenameWest Long Lake      -6.0878       0.6895  -8.829  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.355 on 9719 degrees of freedom
## Multiple R-squared:  0.03953,    Adjusted R-squared:  0.03874
## F-statistic:    50 on 8 and 9719 DF,  p-value: < 2.2e-16
```

13. Is there a significant difference in mean temperature among the lakes? Report your findings.

    Answer: In both the ANOVA and linear models, the p-value is less than the significance level
    (alpha=0.05) and thus we reject the null hypothesis, and conclude that there is strong evidence
    that the mean temperature of at least two different lakes in the month of July is different.

14. Create a graph that depicts temperature by depth, with a separate color for each lake. Add a
    geom_smooth (method = "lm", se = FALSE) for each lake. Make your points 50 % transparent.
    Adjust your y axis limits to go from 0 to 35 degrees. Clean up your graph to make it pretty.

```r
# Scatterplot temperature by depth
ggplot(selected_ntl_nter, aes(x=depth, y=temperature_C)) +
  geom_point(aes(color=lakename)) +
  geom_smooth(method = "lm", se=FALSE, alpha=0.5, color="red") +
  ylim(0,35) +
  ylab(expression("Temperature in"*degree*"C")) +
  xlab("Depth (m)") +
  scale_colour_manual(values = rainbow(9)) +
  ggtitle("Lake temperature by depth")
```

```
## 'geom_smooth()' using formula 'y ~ x'
```

Lake temperature by depth
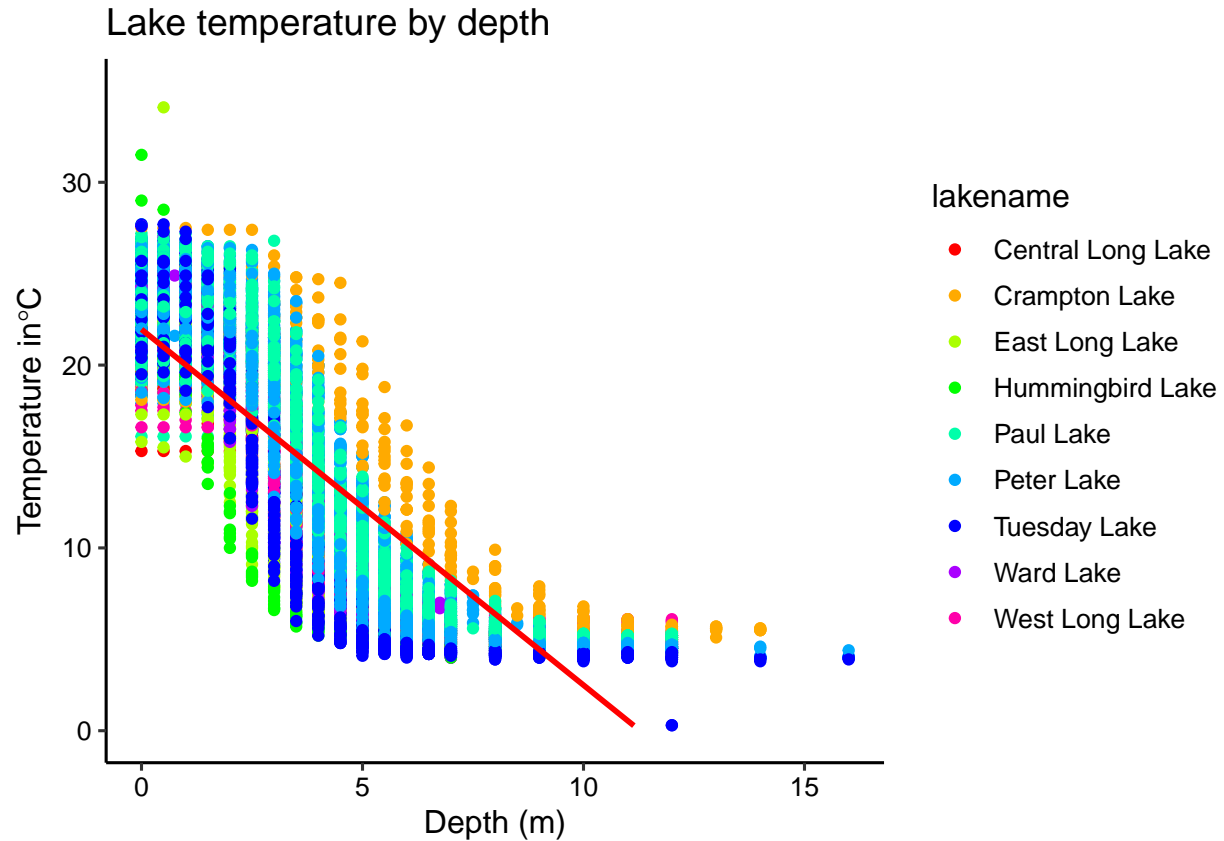
15. Use the Tukey's HSD test to determine which lakes have different means.

```
# Post-hoc test using Tukey Honest Significant Differences
TukeyHSD(lakeTemp_anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = selected_ntl_nter)
##
## $lakename
##                                      diff        lwr        upr      p adj
## Crampton Lake-Central Long Lake    -2.3145195 -4.7031913  0.0741524 0.0661566
## East Long Lake-Central Long Lake   -7.3987410 -9.5449411 -5.2525408 0.0000000
## Hummingbird Lake-Central Long Lake -6.8931304 -9.8184178 -3.9678430 0.0000000
## Paul Lake-Central Long Lake        -3.8521506 -5.9170942 -1.7872070 0.0000003
## Peter Lake-Central Long Lake       -4.3501458 -6.4115874 -2.2887042 0.0000000
## Tuesday Lake-Central Long Lake     -6.5971805 -8.6971605 -4.4972005 0.0000000
## Ward Lake-Central Long Lake        -3.2077856 -6.1330730 -0.2824982 0.0193405
## West Long Lake-Central Long Lake   -6.0877513 -8.2268550 -3.9486475 0.0000000
## East Long Lake-Crampton Lake       -5.0842215 -6.5591700 -3.6092730 0.0000000
## Hummingbird Lake-Crampton Lake     -4.5786109 -7.0538088 -2.1034131 0.0000004
## Paul Lake-Crampton Lake            -1.5376312 -2.8916215 -0.1836408 0.0127491
## Peter Lake-Crampton Lake           -2.0356263 -3.3842699 -0.6869828 0.0000999
## Tuesday Lake-Crampton Lake         -4.2826611 -5.6895065 -2.8758157 0.0000000
```

```
## Ward Lake-Crampton Lake              -0.8932661 -3.3684639  1.5819317 0.9714459
## West Long Lake-Crampton Lake         -3.7732318 -5.2378351 -2.3086285 0.0000000
## Hummingbird Lake-East Long Lake       0.5056106 -1.7364925  2.7477137 0.9988050
## Paul Lake-East Long Lake              3.5465903  2.6900206  4.4031601 0.0000000
## Peter Lake-East Long Lake             3.0485952  2.2005025  3.8966879 0.0000000
## Tuesday Lake-East Long Lake           0.8015604 -0.1363286  1.7394495 0.1657485
## Ward Lake-East Long Lake              4.1909554  1.9488523  6.4330585 0.0000002
## West Long Lake-East Long Lake         1.3109897  0.2885003  2.3334791 0.0022805
## Paul Lake-Hummingbird Lake            3.0409798  0.8765299  5.2054296 0.0004495
## Peter Lake-Hummingbird Lake           2.5429846  0.3818755  4.7040937 0.0080666
## Tuesday Lake-Hummingbird Lake         0.2959499 -1.9019508  2.4938505 0.9999752
## Ward Lake-Hummingbird Lake            3.6853448  0.6889874  6.6817022 0.0043297
## West Long Lake-Hummingbird Lake       0.8053791 -1.4299320  3.0406903 0.9717297
## Peter Lake-Paul Lake                 -0.4979952 -1.1120620  0.1160717 0.2241586
## Tuesday Lake-Paul Lake               -2.7450299 -3.4781416 -2.0119182 0.0000000
## Ward Lake-Paul Lake                   0.6443651 -1.5200848  2.8088149 0.9916978
## West Long Lake-Paul Lake             -2.2356007 -3.0742314 -1.3969699 0.0000000
## Tuesday Lake-Peter Lake              -2.2470347 -2.9702236 -1.5238458 0.0000000
## Ward Lake-Peter Lake                  1.1423602 -1.0187489  3.3034693 0.7827037
## West Long Lake-Peter Lake            -1.7376055 -2.5675759 -0.9076350 0.0000000
## Ward Lake-Tuesday Lake                3.3893950  1.1914943  5.5872956 0.0000609
## West Long Lake-Tuesday Lake           0.5094292 -0.4121051  1.4309636 0.7374387
## West Long Lake-Ward Lake             -2.8799657 -5.1152769 -0.6446546 0.0021080
```

16. From the findings above, which lakes have the same mean temperature, statistically speaking, as Peter Lake? Does any lake have a mean temperature that is statistically distinct from all the other lakes?

Answer: Statistically speaking, Paul lake and Ward lake are the ones that have the same mean temperature as peter lake. There is no lake which has a mean temperature that is statistically distinct from all the other lakes.

17. If we were just looking at Peter Lake and Paul Lake. What's another test we might explore to see whether they have distinct mean temperatures?

Answer: The two-sample t-test is the other test that we can use to compare the two lakes whetehr they have distincit mean temperatures or not.